PATENT
Express Mail No. ET157746906US
Attorney Docket No. LXGN 00102

APPLICATION FOR UNITED STATES OF AMERICA LETTERS PATENT

For

TITLE:

REGION DEFINITION PROCEDURE

AND CREATION OF A REPEAT SEQUENCE FILE

By

Inventor:

Christophe Person

Assignee:

Lexicon Genetics Incorporated

The Woodlands, TX, USA

FIELD OF THE INVENTION

The present invention relates to a system for using a computerized Region Definition procedure in the creation of a Repeat Sequence file.

BACKGROUND OF THE DISCLOSURE

Nucleic acids (DNA and RNA) carry within their structure the hereditary information and are therefore the prime molecules of life. Nucleic acids are found in all living organisms including bacteria, fungi, viruses, plants and animals and they make up the genes within the cell. It is estimated that there are over 100,000 genes within the genome of the human cell. It is of interest to determine the relative abundance of nucleic acids in different cells, tissues and organisms over time under various conditions, treatments and regimes. The nucleic acids code for the amino acids, which are the molecular building blocks of proteins. Proteins are found within the cells of an organism and function to keep the cells alive and responding to it's environment.

Informatics is the study and application of computer and statistical techniques to the management of information. Bioinformatics and computation in biological research have changed dramatically in the last decade. Increasingly, molecular biology is shifting from the laboratory bench to the computer desktop. Today's researchers require advanced quantitative analyses, database comparisons, and computational algorithms to explore the relationships between sequence and phenotype. New observational and data collection techniques have

expanded the capabilities of biological research and are changing the scale and complexity of biological questions that can be productively posed.

The structures of coding and non-coding DNA sequences and amino acid sequences of many organisms have been analyzed, and information concerning those sequences has been recorded in databases accessible via the World Wide Web for common use. Biomedical researchers can gain access to such public domain databases and utilize this information in their own research. Such databases include, for example, GenBank in the U.S., EMBL in Europe, DDBJ at National Gene Institute of Japan, and so on. Genetic information for a number of organisms has also been catalogued in computer databases. For example, genetic databases for organisms such as Escherichia coli, Caenorhabditis elegans, Arabidopsis thaliana, and Homo sapien sapien, are publicly available. At present, however, complete sequence data is available for relatively few species and the ability to manipulate sequence data within and between species and databases is limited.

The new wealth of biological data generated by ongoing genome projects is being used by biologists in combination with newly developed tools for database analysis to ask many questions from molecular interactions to relationships among organisms. Bioinformatics, is contributing to the usefulness of the information generated by the genome projects with the development of methods to search databases quickly, to analyze nucleic acid sequence information, and to predict protein sequence, structure and correlate gene function information from DNA sequence data. Comparisons of multiple sequences can reveal gene functions that are

not evident in any single sequence. Web-based searches of several collections of amino acid sequence motifs can elucidate particular structural or functional elements.

Biological sequence databases, though, contain many repeated and redundant sequences or sequence fragments. These repeated and redundant sequences or sequence fragments have been deposited in the sequence repository databases as many as three or more times. Sequences may be deposited redundantly because often researchers from different laboratories determine the sequences of the same gene or chromosome segment from the same or closely related species. Some identical or closely related sequences have been deposited approximately 10^3 times in the biological sequence databases. Repeated sequences appear naturally in the DNA/RNA and are deposited as part of a whole sequence or fragment. In addition, a variety of experimental protocols contribute to the increase of contamination sequences deposited in databases. Because of such contamination, some chimeric sequences produced from different genes of different species (yeast, bacteria, etc.) may be present.

There is an existing need for a fast-computerized method of identifying and masking repeat and redundant sequences. Redundancies in the currently available DNA/RNA databases render the systematic analysis of similarity or homology between DNA/RNA sequences impractical both in terms of computation and time. Both repeated and redundant sequences present a special problem when searching the public domain and other biological sequence databases for related sequences. If a given Query matches a repeated or redundant sequence, the large number of resulting matches may obscure interesting relationships to other less related but

still informative genes. The conventional bioinformatic algorithms available do not address these problems.

SUMMARY OF THE INVENTION

The disclosure teaches a method for identifying repeated sequences within Redundant Sequence Database Files (RED FILES) via a Region Definition and Transition Identification procedure. Sequences from the RED FILES can be searched and rendered more useful by first identifying repeated sequences within them. Subsequently, identified repeated sequences can be stored in a separate Repeat Sequence Database File (REP FILE) for future identification and masking processes.

One aspect of this invention is a method for identifying a repeat sequence. This method includes selecting a query sequence, comparing the query sequence with other sequences in a redundant file, identifying sequences in the redundant file that contain a similar sequence to a portion of the query sequence, aligning all identified sequences with the similar sequence in the query sequence, designating the right and left endpoints of each identified sequence and any intervening sequences, identifying a position within the query sequence corresponding to each endpoint, defining regions within the query sequence where a region is a sequence between two consecutive positions corresponding to two endpoints, and identifying all regions having at least five sequence matches in the redundant database as repeat sequences.

Another aspect of the invention is a method for constructing a repeat database. This method includes selecting a query sequence, selecting known repeat sequences, adding known

repeat sequences into a repeat sequence database, masking the query with repeat sequences in the repeat sequence database, comparing the masked query sequence with other sequences in a redundant file, identifying sequences in the redundant file that contain a similar sequence to a portion of the query sequence, aligning all identified sequences with the similar sequence in the query sequence, designating the right and left endpoints of each identified sequence and any intervening sequences, identifying a position within the query sequence corresponding to each endpoint, defining regions within the query sequence where a region is a sequence between two consecutive positions corresponding to two endpoints, identifying any two successive regions having a large variance in the number of sequence matches, and adding the sequence within the region of the two successive regions having the highest number of sequence matches into the repeat sequence database.

The foregoing has outlined rather broadly the features and advantages of the present invention in order that the detailed description of the invention that follows may be better understood. Additional features and advantages of the invention will be described hereinafter which form the subject of the claims of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

The novel features which are believed to be characteristic of the invention will be better understood from the following detailed description, in conjunction with the accompanying drawings.

Figure 1. Illustrates a preferred ordering of subsets of Redundant Sequence Database Files (RED FILES).

Figure 2. Illustrates a flow diagram of key steps employed in identifying the repeat sequences used to generate a Repeat Sequence Database File (REP FILE).

Figure 3. Illustrates a pairwise sequence alignment with gaps in the sequences, where the bases of $Q_i$left and $Q_i$right align exactly with $H_i$left and $H_i$right from the Query/Hit pairwise alignment fragment$_i$.

Figure 4A. Illustrates three examples of pairwise alignments where the Hit sequence fragments are lined up in relationship to the original Query Sequence.

Figure 4B. Illustrates how Boundary Regions are defined using a graphical local multiple sequence alignment output with three Hit Sequences.

Figure 5A. Illustrates three examples of pairwise alignments.

Figure 5B. Illustrates how Boundary Regions are defined using a local graphical multiple sequence alignment output with three Hit Sequences.

Figure 6. Illustrates the Transition Point Definition and Repeat Sequence recognition using a graphical multiple sequence alignment with multiple Hit Sequences with and without open areas created during the alignment.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

This disclosure teaches a computerized method of a Region Definition Procedure that increases the efficiency of standard bioinformatics tools and databases. This procedure is designed to enhance the specialized needs of a high-throughput genomics-computing environment by identifying highly repetitive sequences and storing them in a Repeat Sequence Database File (REP FILE). The REP FILE can be used to mask highly repetitive sequences within a Query Sequence before proceeding further with database sequence comparisons.

1. Relevant Terminology

There is some ambiguity in the scientific literature as to the relevant nomenclature, so it is important to define some specific terms within this disclosure. The following bioinformatics terms are used to define concepts throughout the specification. The descriptions are provided to assist in understanding the specification, but are not meant to limit the scope of the invention.

A Repeat Sequence Database File (REP FILE) is composed of sequence blocks that are known to be present in multiple copies in a single genome, etc. (e.g., Alu sequences).

Public Domain Sequence Databases are databases available for use by the public. Typically, such databases are maintained by an entity that is different from the entity creating and maintaining the REP FILE. In the context of this invention, the public domain databases are used primarily to obtain information about the Query Sequences obtained from other sequencing laboratories around the world. Examples of such Public Domain Databases include the GenBank

and dbEST databases maintained by the National Center for Biotechnology Information (NCBI), TIGR database maintained by The Institute of Genomic Research and SwissProt maintained by ExPasy.

Redundant Files (RED FILES) include public domain sequence databases and Independent Sequence Databases that contain redundant sequences. Query Sequences are selected from the RED FILES and generally contain several redundant sequences. Redundant sequences or sequence fragments have been deposited in the sequence repository two or more times. Sequences may be deposited multiple times because researchers from different laboratories determine the sequences of the same gene or chromosome segment from the same or closely related species or because the sequence is a commonly repeated sequence domain within a gene. Some identical or closely related sequences have been deposited approximately 10^3 times in the public domain sequence databases, generating redundancies that are costly in terms of processing and analysis.

Target database(s) are databases of pre-existing sequences to which the Query Sequence will be compared to find the most similar matches (example: UNIQUE and REP FILES).

Database Search Algorithms are mathematical means of identifying similar sequence regions within a Query Sequence when compared to database sequences. BLAST, FASTA, Smith-Waterman are common examples of database search algorithms that can produce a list of pairwise alignments between a Query Sequence and all matching (Hit) sequences in searchable sequence databases.

A Cluster is a group of sequences related to one another by sequence similarity. Clusters are generally formed based upon a specified degree of homology similarity and overlap.

An Algorithm is a mechanical or recursive computational procedure for solving a problem.

A Multiple Sequence Alignment (MSA) is a group of three or more sequences aligned to maximize the registry of identical residues. Global MSA are sequence alignments that require the participation of all sequence residues. For the purpose of this disclosure Local MSA will be used that does not require the participation of all sequence residues in the alignment. MSA is the process of aligning several related sequences, showing the conserved and non-conserved residues across all of the sequences simultaneously. These conserved/non-conserved residues form a pattern that can often be used to retrieve sequences that are distantly related to the original group of sequences. These distant relatives are extremely helpful in understanding the role that the group of sequences plays in the process of life. This can be the alignment of like nucleic acid residues of several genes or the amino acids of a number of protein sequences. The final product of a MSA may contain a gap character, "-", which is used as a spacer so that each sequence has the same number of residues plus gaps in the alignment. A MSA shows the residue juxtaposition across the entire set of sequences; thus showing the conserved and non-conserved residues across all of the sequences simultaneously.

A Scoring Matrix is a table of values used to evaluate the alignment of any two given residues in a sequence comparison. For protein sequences there are two main families of scoring matrices: PAM and BLOSUM.

FASTAlign is Lexicon Genetics' clustering software for the rapid construction of multiple sequence alignments from nucleotide and protein sequences. FASTAlign is a multiple sequence alignment algorithm similar to NCBI's N-align.

BLAST (Basic Local Alignment Search Tool) is a set of database search programs designed to examine sequence databases. BLAST uses a heuristic algorithm which seeks local as opposed to global alignments and is therefore able to detect relationships among sequences which share only isolated regions of similarity (Altschul et al., 1990).

FASTA is a set of sequence comparison programs designed to perform rapid pairwise sequence comparisons. Professor William Pearson of the University of Virginia Department of Biochemistry wrote FASTA (Pearson, William, 1990). The program uses the rapid sequence algorithm described by Lipman and Pearson (1988) and the Smith-Waterman sequence alignment protocol.

The Smith-Waterman Algorithm is a modification of the global alignment method that efficiently identifies the highest scoring sub-region shared by two sequences (Smith and Waterman, 1981, Waterman, M.S., 1989 and Waterman, M.S., 1995). Often homologous

sequences only share similarity in a small sub-region. Global alignments may fail to include such regions of relatedness in an end-to-end optimal alignment.

An Expectation Threshold (ET) is the length of a sequence alignment determined to be necessary to distinguish between evolutionary relationships and chance sequence similarity. The ET is calculated using normalized probability scores. The ET selected will vary based on the amount of error one is willing to accept. For example, an ET of 8 nucleotides can be accepted if one is willing to accept an 8-10% error. If one is only willing to accept a small percentage error, then the ET selected must be a longer nucleotide sequence. Preferably, a minimum ET of 100 nucleotides is selected for determining if a portion of a Query Sequence is a Unique Sequence. However, where a Hit contains a relatively small area having no matching nucleotides in the Query Sequence, an ET of about 30 nucleotides may be selected.

N-Align is a program that NCBI uses to recast the standard bioinformatic database output. The Query/Hit Sequence pairs, identified from database searches, are aligned to the full Query Sequence. This alignment format exists in graphical and text renditions in the NCBI search outputs.

A Sequence Database Search Output consists of a collection of one or more identified pairwise alignments in a Query-Hit Sequence pair that exceeds a designated expectation threshold (ET).

A Pairwise Alignment is an alignment of a part or a whole of two sequences.

Pairwise alignment software is a program used to recast the standard bioinformatics database output. The Query/Hit Sequence pairs, identified from database searches, are aligned to the full Query Sequence. This alignment format exists in graphical and text renditions in many public search outputs.

A Sequence Alignment is a comparison between two or more sequences that attempt to bring into register identical or similar residues held in common by the sequences. It may be necessary to introduce gaps in one sequence relative to another to maximize the number of identical or similar residues in the alignment.

A Hit is when two or more sequences are brought together into register with identical or similar residues that are held in common by those sequences in a pairwise alignment.

The following definitions are used to define molecular biology terms throughout the specification. These definitions are provided to assist in understanding the specification, but are not meant to limit the scope of the invention.

A contig is a group of overlapping DNA segments.

A contig map is a chromosome map showing the locations of those regions of a chromosome where contiguous DNA segments overlap. Contig maps are important because they provide the ability to study a complete, and often large segment of the genome by examining a

series of overlapping clones which then provide an unbroken succession of information about that region.

A Consensus sequence is a nucleotide sequence constructed as an idealized sequence in which each nucleotide position represents that base most often found at that position when many related nucleotide sequences are compared. Variations of mismatch nucleotides compared to consensus sequences may characterize single nucleotide polymorphisms (SNPs) representing the diversity or polymorphism of a particular gene in the population or species.

A Concatamer is a global consensus sequence created by joining end to end overlapping sequence fragments and merging areas of the overlap.

A Gene is the functional and physical unit of heredity passed from parent to offspring. In this disclosure the term gene is intended to mean a sequence of bases of DNA or mRNA bases containing the information to code for a sequence of amino acids that make up a protein.

2. Sequence Query Acquisition and Building a Repeated Sequence File.

Figure 1 and 2 present a preferred embodiment of a fast-computerized method of identifying repeated sequences within the Redundant Sequence Database Files (RED FILES) via a Region Definition and Transition Identification procedure and placing them into a Repeat Sequence File (REP FILE). A more detailed discussion of the steps in this process is described below. Sequences from the RED FILES can be searched and rendered more useful by first

identifying repeated sequences within them. Subsequently, identified repeated sequences can be stored in a separate REP FILE for future identification and masking processes.

As shown in Figure 1 a Query Sequence 104 is selected for a repeated sequence search from an ordered subset of RED FILES 102. For access to the most useful data available in the public domain, this subset of the RED FILES has been ordered by species and by annotation richness. Generally, the first set of Query Sequences to be selected is from the Human mRNA database files in the RED FILES 102. The Human database subset is the most relevant species for medical research and is typically the first database to be searched for repeat sequences. The Human mRNA databases have very rich or excellent annotations. However, depending on the Query sequence, it may be more relevant to use other species sequences. In the following paragraphs, Human can be substituted with any other species, depending on the intents and goals of the user. All annotations associated with the selected Query Sequences will be maintained and stored with the Query Sequence or any subsequently identified fragment thereof.

Mouse mRNA database files, which is a very large database with very good annotations, is generally searched for repeats after the Human mRNA subset has been searched.

The other database subsets, such as the total RNA, Mouse EST and Human EST, are preferably searched in the order of the richness of their annotations and future usefulness in correlating gene function and location information from genomic DNA sequence data. However, if the investigator is interested specifically in the mouse database files, Queries from the mouse RNA database files would be selected first.

As shown in Figure 2 the selected Query Sequence 104 will be tested and masked 205 against the Repeat Sequence Database (REP FILE) 207. The REP FILE is composed of sequences and fragments that are not unique and are known to be present in multiple copies in a single genome (e.g., Alu sequences, *E. coli* sequences, blue script sequences, etc.). These sequences may be present in the selected Query Sequence 104 and must be eliminated or masked before new repeats can be identified. The masked Query Sequence is then tested 209 against the RED FILE subset 211. The RED FILE subset is known to contain repeat and redundant sequences or sequence fragments that have been deposited in the sequence repository two or more times.

The analysis systems, represented by step 205 and 209 in Figure 2 in process flow 200 may use typical programs, such as the Smith-Waterman algorithm (Smith and Waterman, 1981, Waterman, M.S., 1989 and Waterman, M.S., 1995), the BLAST programs (Altschul et al., 1990), or the FASTA program (Pearson, William, 1990, Lipman and Pearson, 1988), or any pairwise sequence alignment program or method to test the Query Sequence.

These programs use rapid sequence alignment algorithms that produce a list of pairwise alignments. A parsing program scans the pairwise alignments produced and accumulates them in a buffer. These pairwise alignments are reduced and contigs are created which are then processed back through the sequence alignment algorithm as a new Query Sequence. This alignment and parsing continues until the Query Sequence alignment process identifies all known-matching sequences in the target databases. Scoring Matrix Programs such as PAM

(M.O. Dayhoff, 1978) or the BLOSUM family (Henikoff and Henikoff, 1992) are used to evaluate the matches of the alignment and Expect Values of Altschul (Altschul et al., 1997) is the method of ranking the scores of the matches. Due to sequence polymorphism, and in the context of several million analyses, the validity of the matches may be re-evaluated by other methods in the context of gene specificity. FASTAlign then recasts the compiled text listings of these pairwise alignments into a graphical rendition.

Boundary Regions (as described below in Example 1) are then defined 213 using the multiple sequence alignments created during the testing phase 209. A Boundary Transition algorithm (as described in Example 2) is then used to identify different transition patterns between the Boundary Regions of sequence hits. These transition patterns are used to detect new repeating sequences. The question is then asked, "Are there new repeat sequence fragments in the Query?" 215. If this region meets the pre-set conditions with no overlapping Hit fragment the answer is YES. Pre-set conditions are requirements that must be met for a region to be considered, such as, minimum length, percent quality of this Query region sequence, etc. A "YES" answer 217 will place the new repeating sequence in the REP FILE 219 and a new Query Sequence is chosen. A "NO" answer 221 signals that there is no new repeating sequence and the negative result is ignored 223 and a new Query Sequence is chosen.

EXAMPLE 1 REGION DEFINITION PROCEDURE

A. Comparison of Query Sequence with Target Database

A Query sequence is compared with sequences in a Target Database such as the REP FILES and a subset of the RED FILES (e.g., the Human mRNA subset). Regions are defined based upon the relative position of the endpoints of the similar database sequence or Hit Sequence to the Query Sequence. Each sequence in the Target Database that matched the sequence of a part or all of the Query Sequence is analyzed separately.

B. Identification of Endpoints on the Query Sequence

As illustrated in Figure 3, the endpoints of the Query Sequence are defined as $Q_i$left 302, the left most absolute position of the Query Sequence or the left endpoint of the Query Sequence, and $Q_i$right 306, the right most absolute position of the Query Sequence or the right endpoint of the Query Sequence. When a similar database sequence in the Target database is identified that matches a part or all of the Query Sequence it is then aligned with the part of the Query Sequence that it is similar to. For example, in Figure 3 the Query Sequence and the similar database sequence (hereinafter referred to as a Hit) are almost identical. Thus, the left most absolute position of the Hit ($H_i$left 304) matches the left most absolute position of the Query Sequence ($Q_i$left 302) where the nucleotide at 302 and the nucleotide at 304 are aligned exactly and represent the left most aligned nucleotide pair. Similarly, the right most absolute position of

the Hit ($H_i$right 308) matches the right most absolute position of the Query Sequence ($Q_i$right 306) where the nucleotide at 306 and the nucleotide at 308 are aligned exactly and represent the right most aligned nucleotide pair. The alignment of these two sequences represents one pairwise alignment.

Figure 4A illustrates the relative positional relationships between three Hit Sequences 402, 404, 406 and the Query Sequence 422. The first pairwise alignment 450 is composed of Hit Sequence 402 and a portion of the Query Sequence 422 between points 408 and 410. The second pairwise alignment 452 is composed of Hit Sequence 404 and a portion of the Query Sequence 422 between points 412 and 414. The third pairwise alignment 454 is composed of Hit Sequence 406 and a portion of the Query Sequence 422 between points 416 and 418. The Hit Sequences in the pairwise alignments are annotated with the nucleotide numbers from the Query Sequence 422 to which they correspond. For example, if the portion of the Query Sequence 422 between points 408 and 410 represents nucleotides 1 to 150, with the first nucleotide at left most end point being number 1, then the Hit Sequence 402 would be annotated to indicate that it matched the portion of the Query Sequence 422 between nucleotides 1 to 150.

C. Graphical Alignment of the Pairwise Alignments

Software programs such as NCBI's N-align or Lexicon Genetics' FASTAlign are used to recast the pairwise alignments into an ordered graphical format where each of the Hit Sequences are displayed below the entire Query Sequence aligned with the portion of the Query Sequence that it is similar to. Figure 4B shows the graphical alignment of three Hit Sequences 402, 404

and 406 with their similar or homologous sequences aligning with matching areas on the Query Sequence 422.

D. Identifying Similar Sequence Regions

The graphical representation of the alignment of each Hit Sequence with their similar or homologous sequences on the Query Sequence 422 and overlap sequence fragments on any other contiguous Hit Sequence is used to determine the Boundary Regions in Figure 4B. The endpoints of each Hit Sequence are visually connected to the Query Sequence 422. For example, Hit Sequence 402 left and right endpoints are connected to the Query Sequence 422 with dashed lines 408 and 410. The endpoints of Hit Sequence 404 have dashed lines 412 and 414 connecting it to the Query Sequence 422. Similarly, Hit Sequence 406 has dashed lines 416 and 418 connecting it to the Query Sequence 422.

Each of the lines that connect an endpoint of a Hit Sequence may intersect other Hit Sequences, if those Hit Sequences contain an overlapping sequence fragment to the initial Hit Sequence. For example, the dashed line 412 connecting the left endpoint of Hit Sequence 404 to the Query Sequence 422 intersects Hit sequence 402 and the dashed line 414 connecting the right endpoint of Hit Sequence 404 to the Query Sequence 422 intersects Hit Sequence 406. Dashed line 418 indicates the right endpoint of the Query Sequence 422 and the right endpoint of Hit Sequence 406.

When lines connecting all of the Hit Sequence endpoints are drawn to the Query Sequence 422 a series of Boundary Regions (hereinafter referred to as Regions) are visualized.

A Region represents the sequence between two consecutive dashed lines connecting Hit Sequence endpoints to other Hit Sequences and the Query Sequence 422. Each Region (R_1 through R_5 in Figure 4B) is identified and annotated to match the nucleotide sequence that it intersects in the initial Query Sequence 422 so that it can be related directly to a physical location on the original Query Sequence 422.

E. Alignment of Several Missing Nucleotides in a Hit Sequence with the Query Sequence.

Any process for relating a plurality of Hit Sequences to a Query Sequence must take into account areas having several contiguous nucleotides that may be missing within the aligned Hit Sequence. Figure 5A illustrates the relationship between Hit Sequences 502/504, 506 and 508/510 and the Query Sequence 530 where Hit Sequences 502/504 and 508/510 contain large open areas that are missing contiguous nucleotides, such areas having about 30 nucleotides or more, when aligned to the Query Sequence 530. These open areas arise during an alignment when there is not a homologous or similar sequence in the database Hit Sequence in relationship to the initial Query sequence 530. It may indicate that a fragment of that gene has been spliced out.

The first pairwise alignment 550 is composed of Hit Sequence 502/504 matching a portion of the Query Sequence 530 between points 509 and 511. The second pairwise alignment 552 is composed of Hit Sequence 506 matching a portion of the Query Sequence 530 between points 513 and 515. The third pairwise alignment 554 is composed of Hit Sequence 508/510 matching a portion of the Query Sequence 530 between points 517 and 519.

Defining Regions in Hit Sequences containing large open areas that are missing continuous nucleotides requires consideration of those open areas when defining Regions. The gap scoring strategy tends to analyze a fragment's score as the gap extends. For this reason smaller fragments tend to score better than their longer gapped fragment counterpart. In the presence of these open areas, lines are drawn from the endpoints of the open areas as well as the endpoints of the Hit Sequences. For example, in Hit Sequence 502/504 (shown in Figure 5B) four lines are drawn that connect endpoints back to the Query Sequence 530. Dashed line 509 connects the left endpoint of the Hit Sequence 502 to the Query Sequence 530, solid line 501 connects the left endpoint of the open area (Region 2, R_2) of Hit Sequence 502 to the Query Sequence 530. Solid line 503 connects the right endpoint of the open area (Region 2, R_2) of Hit Sequence 504 to the Query Sequence 530 and dashed line 511 connects the right endpoint of the Hit Sequence 504 to the Query Sequence 530.

In Hit Sequence 506, solid line 513 and dashed line 515 are drawn from the left and right endpoints of that Hit Sequence 506 to the Query Sequence 530 respectively. The left endpoint of Hit Sequence 506 is a solid line because it overlays the left endpoint 501 of the open area of the Hit Sequence 504. Hit Sequence 508/510, contains an open area (Region 7, R_7) like Hit Sequence 502/504, and has a dashed line 517 connecting the left endpoint of the Hit Sequence 508 to the Query Sequence 530, solid line 505 connecting the left endpoint of its open area (Region 7, R_7) to the Query Sequence 530, solid line 507 connecting the right endpoint of the open area (Region 7, R_7) to the Query Sequence 530, and dashed line 519 connecting the right endpoint of Hit Sequence 510 to the Query Sequence 530.

Endpoint delineation of the Hit Sequences, including any open areas of about 30 nucleotides in length contained therein, is performed with lines drawn back to the Query Sequence 530. This process visualizes the Regions (R_1 through R_8). Each Region is defined on its right and left extremities by an endpoint line.

Whenever a defined Region represents a very small number of nucleotides, as for example less than about 5-10 nucleotides, those Regions can be ignored as an independent Region and incorporated into the next Region to prevent dilution of the significance of the delineated Regions.

EXAMPLE 2 REGION TRANSITION AND REPEAT SEQUENCE IDENTIFICATION

Once the Regions have been defined for all Hit Sequences (as shown in Figure 6) in relation to the Query Sequence 622, the number of sequences, sequence fragments or open areas that are encompassed in each Region are counted. In Figure 6, Region 1 (R_1) encompasses 12 matching sequences or sequence fragments 601-612; R_2 encompasses 2 matching sequence fragments 612, 613 which are each less than about 5 nucleotides long. Region 2 is ignored as a separate region because these fragments are so short and it is included within Region 3. Region 3 (R_3) encompasses 4 matching sequence fragments 612, 613, 614, 615; and R_4 encompasses 5 sequence fragments 612, 613, 614, 615, 616. R_5 also encompasses 5 matching sequence fragments 612, 613, 614, 615, 616; and R_6 encompasses 3 matching sequence fragments, 615, 616, 619 and 1 open area with missing aligned nucleotides 614. R_7 encompasses 4 matching

sequence fragments 614, 617, 618, 619; and R_8 encompasses 3 matching sequence fragments 614, 617, 618. R_9 encompasses 2 matching sequence fragments 617, 618; and R_{10} encompasses 1 matching sequence fragment 617.

A Transition Point is defined as two successive Regions having an unexpectedly high variation in the number of sequences, sequence fragments or gaps encompassed within the Regions. In Figure 6, a Transition Point is found between R_1 and R_3. This determination is made because R_1 had 12 matched sequences or sequence fragments and R_3, successive to R_1 since R_2 was ignored, had only 4 sequence fragments encompassed within it. An alteration in the number of sequence matches within two successive Regions of about 5 or more identifies a Transition Point. At each Transition Point, the Region of the two successive Regions having the higher number of matches is defined as a Repeat. All novel Repeats are identified, stored and added into the REP FILE. In Figure 6, R_1 would be defined as a Repeat and added into the REP FILE.

REFERENCES

Altschul, Stephen F., Gish, W., Miller, W., Myers, W. W. and Lipman, David J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215:403-410.

Altschul, Stephen F., Madden, Thomas L., Schaffer, Alejandro A., Zhang, Jinghui, Zhang, Zheng, Webb Miller, and Lipman, David J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 25:3389-3402.

Dayhoff, M.O. (1978.), in *Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 3, 229-249, National Biomedical Research Foundation, Washington, D.C., M.O. Dayhoff, ed.

Feng D. F., Johnson, M.S. and Doolittle, R.F. (1984-85). Aligning amino acid sequences: comparison of commonly used methods. J Mol Evol. 21(2):112-25.

Henikoff S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. Nov 15; 89(22):10915-9.

Karlin, S. and Ghandour, G. (1985). Multiple-alphabet amino acid sequence comparisons of the immunoglobulin kappa-chain constant domain. Proc Natl Acad Sci U S A. Dec; 82(24):8597-601.

Lipman, David J. and Pearson, W.R. (1985). Rapid and sensitive similarity searches. *Science* 227:1435-1441.

Pearson, W. and Lipman, David (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* 85:2444-2448.

Pearson, W. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. in *Methods in Enzymology* 183, Doolittle, R. ed. cf. pp. 75-85.

Smith, T.F. and Waterman, M.S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147; 195-197.

Waterman, M.S. (1989). Sequence Alignments in *Mathematical Methods for DNA Sequences*, Waterman, M.S. ed. pp. 53-92. CRC Press, Boca Raton.

Waterman, M.S. (1995). Dynamic Programming Alignment of Two Sequences, in *Introduction to Computational Biology: Maps, Sequences and Genomes*. pp. 183-232, Chapman and Hall, New York.

All patents and publications mentioned in this specification are indicative of the level of skill of those of knowledge in the art to which the invention pertains. All patents and publications referred to in this application are incorporated herein by reference to the same extent as if each was specifically indicated as being incorporated by reference and to the extent that they provide materials and methods not specifically shown.